



## METHOD ARTICLE

# Prediction of cell position using single-cell transcriptomic data: an iterative procedure [version 1; peer review: awaiting peer review]

Andrés M. Alonso <sup>1,2</sup>, Alejandra Carrea<sup>1</sup>, Luis Diambra <sup>1</sup>

<sup>1</sup>CREG-CONICET, Universidad Nacional de La Plata, La Plata, Buenos Aires, 1900, Argentina

<sup>2</sup>INTech-CONICET, Universidad Nacional de San Martín, Chascomus, Buenos Aires, Argentina

**v1** First published: 18 Oct 2019, 8:1775 (<https://doi.org/10.12688/f1000research.20715.1>)

Latest published: 18 Oct 2019, 8:1775 (<https://doi.org/10.12688/f1000research.20715.1>)

## Open Peer Review

**Reviewer Status** Awaiting Peer Review

Any reports and responses or comments on the article can be found at the end of the article.

## Abstract

Single-cell sequencing reveals cellular heterogeneity but not cell localization. However, by combining single-cell transcriptomic data with a reference atlas of a small set of genes, it would be possible to predict the position of individual cells and reconstruct the spatial expression profile of thousands of genes reported in the single-cell study. To develop new algorithms for this purpose, the Dialogue for Reverse Engineering Assessments and Methods (DREAM) consortium organized a crowd-sourced competition known as DREAM Single Cell Transcriptomics Challenge (SCTC). In the spirit of this framework, we describe here the proposed procedures for adequate reference genes selection, and an iterative procedure to predict spatial expression profile of other genes.

## Keywords

Single-Cell RNA sequencing, Drosophila Embryo, Gene expression Patterns, DREAM Challenge



This article is included in the **DREAM Challenges** gateway.

**Corresponding author:** Luis Diambra ([ldiambra@gmail.com](mailto:ldiambra@gmail.com))

**Author roles:** **Alonso AM:** Conceptualization, Formal Analysis, Writing – Original Draft Preparation; **Carrea A:** Conceptualization, Formal Analysis, Writing – Review & Editing; **Diambra L:** Conceptualization, Formal Analysis, Supervision, Writing – Original Draft Preparation

**Competing interests:** No competing interests were disclosed.

**Grant information:** The author(s) declared that no grants were involved in supporting this work.

**Copyright:** © 2019 Alonso AM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Alonso AM, Carrea A and Diambra L. **Prediction of cell position using single-cell transcriptomic data: an iterative procedure [version 1; peer review: awaiting peer review]** F1000Research 2019, 8:1775 (<https://doi.org/10.12688/f1000research.20715.1>)

**First published:** 18 Oct 2019, 8:1775 (<https://doi.org/10.12688/f1000research.20715.1>)

## Introduction

Multicellular organisms show throughout their development a crescent cellular heterogeneity, distributed and organized in different organs and tissues. This spatial heterogeneity has been explored using different techniques, such as immunohistochemistry and single-molecule fluorescence *in situ* hybridization (FISH)<sup>1</sup>. These approaches allow quantification of gene expression in many cells but, unfortunately, these techniques can currently only be assayed to a small number of genes. The selection of these genes introduce a bias that limits the power of these studies. With the advent of emergent methods in genomics, it has become possible to assess the transcriptomic profile of complex tissues with unprecedented resolution, thereby allowing insights into complex processes such as: differentiation trajectories, cell fate decisions, and spatial relationships. In this sense, high-throughput single-cell RNA-seq (sc-RNA-seq) is becoming an established experimental technique<sup>2</sup>. The protocol of this technique includes the initial step of sample collection, during which solid tissue dissociation results in single cells. Separating cells from their native context results in the loss of spatial information. However, investigating the molecular composition of individual cells in the context of spatial location is important, especially when studying primary cancer cells<sup>3</sup>. However, some progress has been made to overcome limitations of spatial information loss associated to this technique. Computational methods, based on Principal Component Analysis, are able to partially recover the spatial structure of gene expression patterns<sup>4</sup>. More recently several computational techniques, coupled to *in situ* RNA patterns facilitate this reconstruction with better resolution<sup>5-7</sup>.

In order to catalyze research on computational methods for the spatial reconstruction of single-cell gene expression data, a crowd-sourced competition was designed by the DREAM Consortium in collaboration with Nikos Karaiskos and Nikolaus Rajewsky from Max Delbrück Institute. Using sc-RNA-seq data from Rajewsky Lab, published in 7, and the expression patterns of driver genes as an expression reference atlas, three main subchallenges were designed. The particular aim was to predict the position of 1297 cells in the 3039 *Drosophila melanogaster* embryonic locations, or bins, for one half of an embryo in stage 6 (pre-gastrulation), based on scRNAseq data. The prediction of the 1297 cell positions must be done using a limited number of genes selected from a pool of 84 expression patterns used as a reference atlas. In subchallenge 1, the prediction must be performed using 60 driver genes out of 84 genes; subchallenge 2, using a subset of any expression patterns from 40 genes out of the 84; subchallenge 3, using a subset of any expression patterns from only 20 driver genes. The selection of the subset of genes used for the prediction poses an additional and interesting problem.

## Methods

### DREAM challenge data

Expression patterns used as a reference atlas correspond to 84 driver genes obtained from *in situ* hybridization experiments; the data correspond to The Berkeley Drosophila Transcription Network Project (BDTNP)<sup>8</sup>. This gene expression data set is listed in the file `bdtnp.csv` at DVEX server. One half of the

*Drosophila* embryo has 3039 cells locations, each location is specified by three coordinates ( $x$ ,  $y$  and  $z$ ) (`geometry.txt` at DVEX). Thus, the reference database consists of an expression matrix of 84 genes (columns) quantified across the 3039 embryonic locations (rows). These data were binarized following<sup>7</sup>, sorted in the same order of cell location, and listed in an additional file (`binarized_bdtnp.csv` at DVEX server). The single-cell RNA sequencing data is provided as a matrix with 8924 genes as rows and 1297 cells as columns. These data are divided by the total number of counts for that cell, in this step a pseudocount is added. The normalized values are obtained by taking the logarithm of the total counts. The normalized values are also binarized, i.e., a given gene is ON (OFF) if the normalized values are above (below) of a quantile value. Based on a distance minimization criterion, the quantile value was chosen as 0.23. The short sequences for each of the 1297 cells in the raw and normalized data are the barcodes of individual cells. Both normalized, as well as binarized, data were provided by the DREAM Challenge.

### Selection of the gene sets

In order to select the gene sets to be used in each subchallenge, we take into account two criteria:

- (i) Genes that have complementary expression patterns across the single-cell population.
- (ii) Genes with expression levels broadly distributed across the single-cell population.

To accomplish these criteria, we first perform an agglomerative clustering procedure over the expression matrix comprising the 84 genes (the same genes as the available in the *in situ* expression data) over the 1297 cells. We cluster genes with similar expression profiles across the cells, by means of using the Euclidean distance over the normalized gene expression levels, and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) as a linkage method. Then, we cut the dendrogram tree into 20, 40, or 60 groups depending on the subchallenge. This step allows us to identify genes, or cluster of genes, with complementary expression patterns; however, we need to select only one gene per cluster. This selection is performed based on the criterion of the broadest distribution. To this end, for each gene within a given cluster we compute the frequency distribution  $p_i$ , where  $p_i$  denotes the frequency of occurrence of expression levels within the bin  $i$ . Here, we set the bin size equal to 0.125. After that, we compute the associated entropy  $S = -\sum_i^n p_i \ln p_i$ . Then we select the gene with the greatest entropy in each cluster, i.e., the gene within the cluster with the broadest expression distribution across the single-cell population. This selection procedure is performed with the R script named `preprocessing.r`, which uses the function `selgen.R`, both available at Zenodo (see *Data availability*). To assess this method for the gene selection, we compare the prediction performance obtained with the set of 20 genes selected in this way with the results obtained with different sets of genes sampled at random. For comparison we consider the Mathews correlation coefficient (MCC) between the 1297 cells and the 3039 bins, the ten better scored bins are selected as putative position for each cell. As the true positions of the cells is not

available, we take the bin with the highest MCC, obtained with the set that include all 84 genes, as the bin associated with the true position. Thus, we count cells with ten best scores containing the true position as cells whose positions are well predicted. The percentage of the well-predicted positions will be our measure of the performance. **Figure 1** depicts the histogram of percentage of cells with well-predicted positions, obtained with 200 sets of 20 randomly selected genes. In all cases this percentage is quite lower than that obtained with 20 genes selected as indicated above, which is 33.46%.

We use this procedure to select an additional set of 100 genes from the 8924 genes measured by single-cell technique, but excluding the genes from the 84 reference gene set. These 100 genes will be used in further steps during the iterative procedure, and will be denoted as the out group set hereafter. The 20, 40 and 60 selected genes used for each cell location prediction task were listed in Table S1 (see *Extended data*); we also include the out group set.

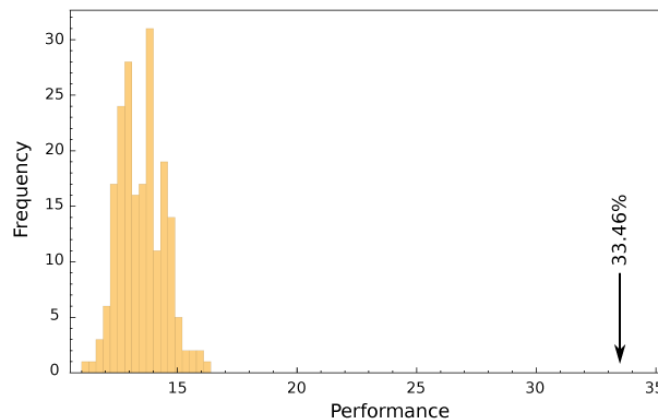
### Scoring functions

In order to predict the position of a given single cell, we use a score approach based on two similarity measures between the

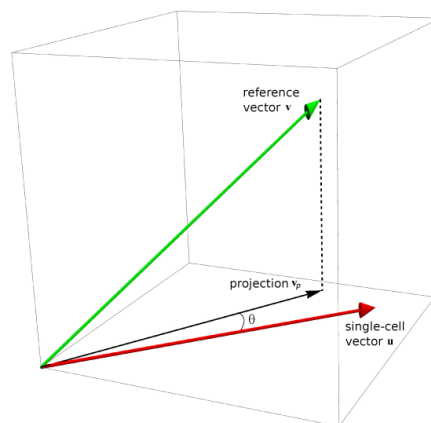
sc-RNA-seq data, and the reference atlas. One of these measures is the Matthews correlation coefficient (MCC) computed between the binarized expression profiles, as proposed in 7. The MCC will be used in the initial step to assign putative bin positions for each single cell and then to predict the spatial expression profile of the outgroup genes. The other measure is the overlap between the normalized expression vector of single cells, and the projected vector corresponding to the predicted spatial expression profile. This vectorial space corresponds to the one spanned by the outgroup genes only. The overlap is defined by:  $\cos(\theta) = \frac{\mathbf{u} \cdot \mathbf{v}_p}{\|\mathbf{u}\| \|\mathbf{v}_p\|}$ , where  $\mathbf{u}$  is the profile vector of the single cell and  $\mathbf{v}_p$  is the vector obtained by projecting the profile vector of the predicted profile on the subspace spanned by the non-null components of the profile vector  $\mathbf{u}$ , as illustrated in **Figure 2**. The scoring functions are performed by the R script named `functions.r`, available at Zenodo (see *Data availability*).

### Results

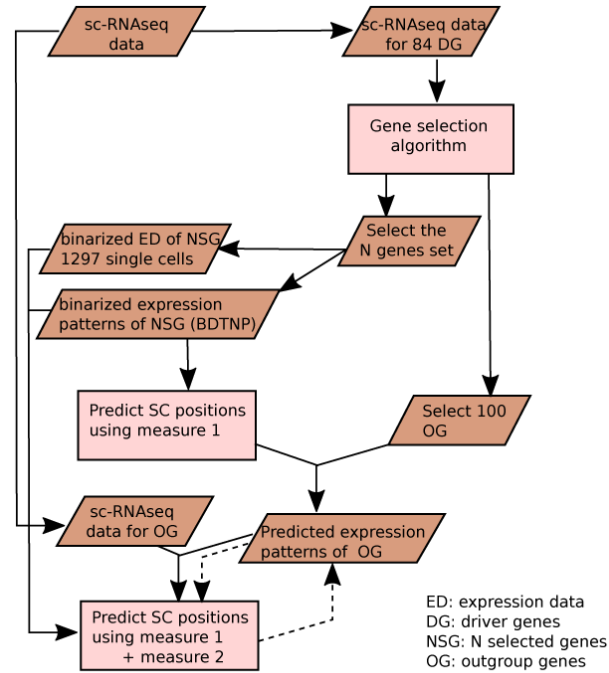
The proposed procedure is schematically illustrated in **Figure 3**. In the first step we select the set of  $N$  genes from the 84 driver genes to be used in the prediction using the method described in



**Figure 1.** Histogram of the performance obtained with random selection of 20 genes (yellow). The performance obtained with the set of 20 genes selected by the proposed method is indicated with a black arrow.



**Figure 2.** Low-dimensional representation of the angle between expression vector  $\mathbf{u}$ , and the projected expression vector  $\mathbf{v}_p$ .



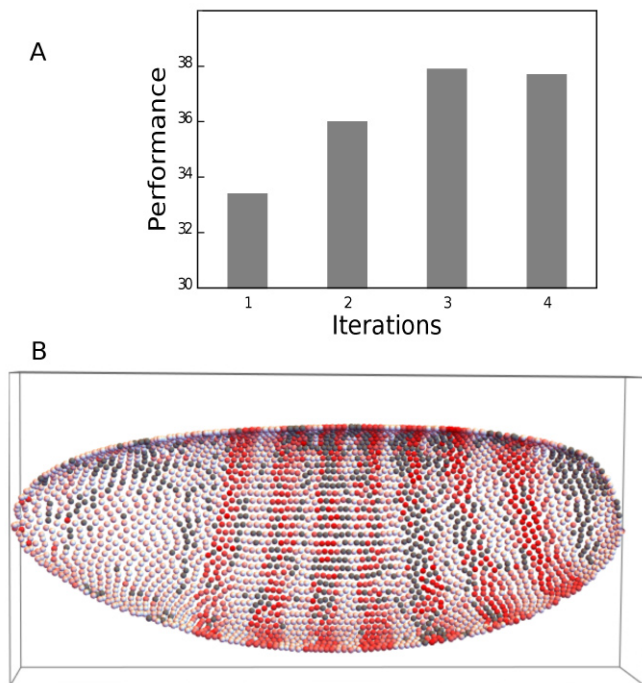
**Figure 3.** Flow diagram of the proposed method.

*Selection of the gene sets* section. We also select an additional 100 genes (outgroup genes) from all genes measured in the sc-RNAseq experiment, but excluding the driver genes. The name of the genes used are listed in *Extended data: Table S1*. Then, using the binarized expression data of the selected genes we compute the MCC (measure 1) for each binarized single-cells vector against the 3039 binarized vectors associated with each positional bin of the reference atlas (BDTNP). By means of the MCC-based score we predict the single cell positions and build the putative expression patterns of the outgroup set of genes. In this sense, the expression level of gene  $g$  at the bin position  $i$  is given by the weighted average of the normalized gene expression across  $N$  putative positions corresponding to that bin, being the weight proportional to the associated MCC. Mathematically,  $e_{ij}^g = \sum_j c_{ij} e_{ij}^g$ , where  $c_{ij}$  are the MCC-based scores of the single cell  $j$  against position  $i$  and  $e_{ij}^g$  are the expression levels of gene  $g$  recorded in the individual cells  $j$ . The asterisk in the summation indicates that the first better scored  $N$  cells are included. The predicted expression patterns computed in this manner are used to compute the overlap (measure 2) with the corresponding expression level of each one of the 1297 single cells. Finally, using the measure 1 and measure 2 we compute a composed score  $S$ , defined as  $S = w_1 * c + w_2 * o$ , where  $c$  is MCC-based score,  $o$  is overlap-based score and  $w_1$  and  $w_2$  are the respective weights. The score  $S$  is used to predict positions and improve the predicted expression patterns of the outgroup set of genes in each iteration. The last two steps are repeated (2 or 3 times), as indicated in **Figure 3** by dashed arrows.

The above scheme was applied to the Sub-challenges with 20, 40 and 60 genes using different values of the weights. In the first example we apply the procedure to the Sub-challenge 3 and

using the 20 genes we compute the MCC for every cell-bin combination. The first iteration of this scoring procedure leads to a performance of 33.5% in assigning the putative positions to each single cell. By means of using the 20 highest coefficients, we predict the expression patterns of the outgroup set of genes. Then, we compute a scoring measure composed of two terms: the MCC computed in the first step (with a weight  $w_1 = 0.7$ ), and the previously-defined overlap between the expression vector of each single cell and the projected expression vector of the reference atlas, being both vectors composed of the 100 outgroup genes (with a weight  $w_2 = 0.3$ ). The score combining both measures is then used to predict the positions of each single cell, which leads to a performance of 36% in the second iteration and 38% in the third iteration. Further iteration steps does not produce any additional improvement. **Figure 4A** depicts the performance evolution of the procedure using this gene set.

To select the set of 60 genes to be used in subchallenge 1, from the 84 genes available in the reference atlas, we perform the above mentioned agglomerative clustering procedure. Then, the 60 genes with the greatest entropy within each cluster are selected. The names of the resulting genes were listed in the first column of *Extended data: Table S1*. As a first step, we compute the MCC for each binarized single-cell vector, and the corresponding 3039 binarized vectors associated with each positional bin of the reference atlas. By means of using the 20 highest MCC for each cell ( $N = 20$ ), we compute the putative expression patterns of the outgroup set of genes. In this case the used scoring measure was composed by MCC with a weight of 0.90, and the overlap, with a weight of 0.10, of the single-cell expression profiles and the 3039 positions of the predicted expression patterns obtained in the previous step.



**Figure 4. Predicted expression pattern of the *ftz* gene obtained with 60 genes after two iteration steps.** The expression level of each nuclei is given in white-red scale. Gray nuclei correspond to positional bins without prediction.

After two iterations the performance obtained was 95.4%, **Figure 4B** shows the predicted expression pattern of the *ftz* gene obtained using the set of 60 genes. The same procedure was used to predict the positions of single cells by considering a set of 40 genes. Again, these genes were selected as described in Methods. The names of the resulting genes were listed in the second column of *Extended data*: Table S1. In this case the performance obtained reaches 71.4%.

## Discussion

We present three innovations that could represent improvements in regard to the original proposal<sup>7</sup>. One of these innovations is the method for selecting the set of genes to be used as reference in the cell positions prediction task. This set of genes is a good starting point in the presented strategy for position prediction, although we have not explored this method in depth. For example, the Jaccard distance could be used in the clustering procedure instead of the Euclidean distance. We noticed that MCC can overestimate false negatives due to the

fact that sc-RNA-seq are not able to record expression of many genes. This results in profiles with many zeros, even in cases of moderate expression levels. For that reason, our second proposed innovation is an alternative way to make the comparison between profiles, as we used in subchallenge 3. Last but not least, the third innovation is the iterative procedure which improves the performance of any of the alternative strategies presented here. In addition, we noticed that the iterative procedure does not necessarily converge to the correct solution, maybe due to error propagation on the predicted patterns.

## Data availability

### Underlying data

Challenge documentation, including the detailed description of the Challenge design, overall results, scoring scripts, can be found at: <http://www.synapse.org/#!Synapse:syn15665609/wiki>. Data for this Challenge can be downloaded from <http://shiny.mdc-berlin.de/DVEX/>.

Zenodo: Prediction of cell position using single-cell transcriptomic data: an iterative procedure, <https://doi.org/10.5281/zenodo.3470061><sup>9</sup>.

This project contains code and documentation underlying the methods.

Data are available under the terms of the **Creative Commons Zero “No rights reserved” data waiver** (CC0 1.0 Public domain dedication).

### Extended data

Zenodo: Prediction of cell position using single-cell transcriptomic data: an iterative procedure, <https://doi.org/10.5281/zenodo.3470061>.

This project contains the following extended data:

- **Table S1: Selected genes:** first, second and third columns list the name of genes used in the Subchallenges 1, 2 and 3, respectively. The last column lists the names of the outgroup genes.

Data are available under the terms of the **Creative Commons Zero “No rights reserved” data waiver** (CC0 1.0 Public domain dedication).

## Acknowledgements

We are grateful to the DREAM SCTC for curation of the challenge and the evaluation of the models.

## References

1. Battich N, Stoeger T, Pelkmans L: **Image-based transcriptomics in thousands of single human cells at single-molecule resolution.** *Nat Methods*. 2013; **10**(11): 1127–33. [PubMed Abstract](#) | [Publisher Full Text](#)
2. Shapiro E, Biezuner T, Linnarsson S: **Single-cell sequencing-based technologies will revolutionize whole-organism science.** *Nat Rev Genet*. 2013; **14**(9): 618–30. [PubMed Abstract](#) | [Publisher Full Text](#)
3. Sierant MC, Choi J: **Single-Cell Sequencing in Cancer: Recent**

**Applications to Immunogenomics and Multi-omics Tools.** *Genomics Inform.* 2018; **16**: e17.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

4. Durruthy-Durruthy R, Gottlieb A, Hartman BH, *et al.*: **Reconstruction of the mouse otocyst and early neuroblast lineage at single-cell resolution.** *Cell.* 2014; **157**(4): 964–78.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
5. Achim K, Pettit JB, Saraiva LR, *et al.*: **High-throughput spatial mapping of single-cell RNA-seq data to tissue of origin.** *Nat Biotechnol.* 2015; **33**(5): 503–9.  
[PubMed Abstract](#) | [Publisher Full Text](#)
6. Satija R, Farrell JA, Gennert D, *et al.*: **Spatial reconstruction of single-cell gene**

**expression data.** *Nat Biotechnol.* 2015; **33**(5): 495–502.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

7. Karaïskos N, Wahle P, Alles J, *et al.*: **The *Drosophila* embryo at single-cell transcriptome resolution.** *Science.* 2017; **358**(6360): 194–199.  
[PubMed Abstract](#) | [Publisher Full Text](#)
8. Fowlkes CC, Hendriks CL, Keränen SV, *et al.*: **A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm.** *Cell.* 2008; **133**(2): 364–74.  
[PubMed Abstract](#) | [Publisher Full Text](#)
9. Alonso A, Carrea A, Diambra L: **Prediction of cell position using single-cell transcriptomic data: an iterative procedure.** 2019.  
<http://www.doi.org/10.5281/zenodo.3470061>

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**